Introduction to Artificial Intelligence, Fall & Winter 2022
College of Computer Science, Zhejiang University
Problem Set 3: Machine Learning (I)

丁尧相

2022 年 12 月 20 日

Problem 1. 搜索并学习 kd 树数据结构，并了解其在 k 近邻中的应用。（请自行完成，无需在作业中回答）

Problem 2. (***) 在中心为原点的 d 维单位球（半径为 1）内用均匀分布采样 N 个数据点，设至少有一个数据点落入以原点为中心，半径为 r 的单位球的概率不低于 0.9，此时半径 r 至少为多大?

备注：从结论中可以看出，当维数 d 升高时，d 维球上的点越来越倾向于分布在球的外壳附近，从而验证了维数灾难 (curse of dimensionality)。

Problem 3. 请回答下面的问题：

1. 请给出 Lec9 幻灯片第 46 页 bias&variance trade-off 的推导过程。
2. 请推导出 Lec9 幻灯片第 49 页 ridge regression 的解析解。

Problem 4. 在 Lec10 幻灯片的第 32 页，我们介绍了 Soft-Margin SVM 的 primal problem（原问题），试参考前面介绍的线性可分 SVM，推导出其对应的 dual problem（对偶问题）。

Problem 5. (***) 在介绍 AdaBoost 时，我们引入了对 SVM 中 margin 定义的推广。即对于一个二分类器 $f(\mathbf{x}) : \mathcal{X} \to [-1, +1]$，其在数据点 $(\mathbf{x}, y)$ 上的 margin 定义为 $yf(\mathbf{x})$。试回答下面的问题：

1. 若定义多分类器 $f(\mathbf{x}) : \mathcal{X} \to \{1, 2, \ldots, K\}$，即分类器用来预测样本 $\mathbf{x}$ 属于 K 个类中的哪一个，你能否给出对应的 margin 的的定义?
2. 理论上 AdaBoost 可以使用任何分类器作为其基学习器。在你看来，普通线性 SVM，以及使用利用核方法进行变换后的 SVM，它们作为 AdaBoost 的基学习器是否合适？为什么?